On Particle Gibbs Sampling

Sumeetpal S. Singh
CAMBRIDGE UNIVERSITY ENGINEERING DEPARTMENT

joint work: N. Chopin (CREST-ENSAE)

MCMSKI
Charmonix 7 January 2014
Running example (Yu & Meng 2011):

\[ X_0 \sim \mathcal{N}(\mu, \sigma^2), \quad X_{t+1} - \mu = \rho(X_t - \mu) + \sigma \epsilon_t, \quad \epsilon_t \sim \text{i.i.d. } \mathcal{N}(0, 1) \]

\[ Y_t | X_t = x_t \sim \text{Poisson}(e^{x_t}) \]
Running example (Yu & Meng 2011):

\[ X_0 \sim N(\mu, \sigma^2), \quad X_{t+1} - \mu = \rho(X_t - \mu) + \sigma \epsilon_t, \quad \epsilon_t \sim \text{i.i.d. } N(0, 1) \]

\[ Y_t|X_t = x_t \sim \text{Poisson}(e^{x_t}) \]

In general:

\[ X_t|X_{t-1} = x_{t-1} \sim m_\theta(x_{t-1}, x_t)dx_t \]

\[ Y_t|X_t = x_t \sim g_\theta(x_t, y_t)dy_t \]
The posterior: \[ p(\theta, x_{0:T}|y_{0:T}), \quad \theta = (\mu, \rho, \sigma) \]
Inference Objective

The posterior: \[ p(\theta, x_0:T | y_0:T) \], \[ \theta = (\mu, \rho, \sigma) \]

The Gibbs sampler: \( (\theta, x_0:T) \rightarrow (\theta', x_0':T) \)

\[ \sigma'| (x_0:T, \mu, \rho) \sim \text{Gamma} (\cdots) \]

\[ \vdots \]

\[ \mu'| (x_0:T, \sigma', \rho') \sim \text{Normal} (\cdots) \]

\[ x_0':T | (\sigma', \mu', \rho') \]
The posterior: \( p(\theta, x_{0:T}|y_{0:T}) \), \( \theta = (\mu, \rho, \sigma) \)

The Gibbs sampler: \((\theta, x_{0:T}) \rightarrow (\theta', x'_{0:T})\)

\[ \sigma'|(x_{0:T}, \mu, \rho) \sim \text{Gamma}(\cdots) \]

\[ \vdots \]

\[ \mu'|(x_{0:T}, \sigma', \rho') \sim \text{Normal}(\cdots) \]

\[ x'_{0:T}|(\sigma', \mu', \rho') \]

One at a time: \( x_i|(x'_{0:i-1}, x_{i+1:T}, \sigma', \mu', \rho') \)
Particle Gibbs kernel (Andrieu, Doucet and Holenstein, 2010):

Replace Gibbs step for $x_{0:T}$ with

$$X_{0:T}'| (\theta', x_{0:T}) \sim P_{\theta', N}(x_{0:T}, dx_{0:T}')$$
Particle Gibbs kernel (Andrieu, Doucet and Holenstein, 2010):

Replace Gibbs step for $x_{0:T}$ with

$$X'_{0:T}|(\theta', x_{0:T}) \sim P_{\theta', N}(x_{0:T}, dx'_{0:T})$$

- Invariant measure $p(x_{0:T}|\theta', y_{0:T})$
- The kernel is a randomly chosen output of a particle filter, $N$ particles, one fixed to $x_{0:T}$
- Meant to emulate the Gibbs step (change the whole trajectory)
- Typically:

$$x'_{0:T} \neq x_{0:T} \text{ but } x'_i = x_i \text{ for small } i.$$

Increasing $N$ fixes this.
Running Particle Gibbs

Sampling $p(\theta, x_{0:399}|y_{0:399})$

Statistic: counting proportion $x'_i \neq x_i$ for $i = 0, \ldots, 399$
Sampling $p(\theta, x_{0:400} \mid y_{0:400})$

Statistic: autocorrelation $X_0, X_{399}$ (200 particles)
Particle filter

Input: \( \{ x^i_{0:t} \}_{i=1}^N \approx p(x_{0:t} | \theta, y_{0:t}) \)

Output: \( \{ x^i_{0:t+1} \}_{i=1}^N \approx p(x_{0:t+1} | \theta, y_{0:t+1}) \)
Particle filter

Input: \( \{ x_{0:t}^i \}_{i=1}^N \approx p(x_{0:t} | \theta, y_{0:t}) \)

Output: \( \{ x_{0:t+1}^i \}_{i=1}^N \approx p(x_{0:t+1} | \theta, y_{0:t+1}) \)

- Append: \((x_{0:t}^i, X_{t+1}^i)\) where \(X_{t+1}^i \sim m_\theta(x_t^i, dx_{t+1})\)
- Weight: \(x_{0:t+1}^i\) gets weight \(w_{t+1}^i = g_\theta(x_{t+1}^i, y_{t+1})\)

Output: Approximation of \( p(x_{0:t+1} | \theta, y_{0:t+1}) \) are \( N \) independent samples from

\[
\sum_{i=1}^N w_{t+1}^i \delta_{x_{0:t+1}^i} (dx_{0:t+1}')
\]
Sampling from $P_{\theta,N}(x_{0:T}^*, dx_{0:T})$

Idea: $N$ particle system but force $x_{0:T}^1 = x_{0:T}^*$
Sampling from $P_{\theta,N}(x_{0:T}^*, dx_{0:T})$

Idea: $N$ particle system but force $x_{0:T}^1 = x_{0:T}^*$

Intermediate step $t$

Input: $\{x^i_{0:t}\}_{i=1}^N$ with $x^1_{0:t} = x_{0:t}^*$

Output: $\{x^i_{0:t+1}\}_{i=1}^N$ with $x^1_{0:t+1} = x_{0:t+1}^*$
Sampling from $P_{\theta,N}(x_0^*, dx_0^T)$

Idea: $N$ particle system but force $x_0^1 = x_0^*$

**Intermediate step $t$**

Input: $\{x_{0:t}^i\}_{i=1}^N$ with $x_0^1 = x_0^*$

Output: $\{x_{0:t+1}^i\}_{i=1}^N$ with $x_0^1 = x_0^*$

- Append particles $i = 2, \ldots, N$:
  
  $$(x_0:t, X_{t+1}^i) \quad \text{where} \quad X_{t+1}^i \sim m_\theta(x_t^i, dx_{t+1})$$

- Weight: $x_{0:t+1}^i$ gets weight $w_{t+1}^i = g_\theta(x_{t+1}^i, y_{t+1})$
Idea: $N$ particle system but force $x_1^0:T = x^*_0:T$

**Intermediate step $t$**

Input: $\{x^i_{0:t}\}_{i=1}^N$ with $x^1_{0:t} = x^*_0:t$

Output: $\{x^i_{0:t+1}\}_{i=1}^N$ with $x^1_{0:t+1} = x^*_0:t+1$

- Append particles $i = 2, \ldots, N$:

  $$(x^i_0:t, X^i_{t+1}) \text{ where } X^i_{t+1} \sim m_\theta(x^i_t, dx_{t+1})$$

- Weight: $x^i_{0:t+1}$ gets weight $w^i_{t+1} = g_\theta(x^i_{t+1}, y_{t+1})$

Output $x^1_{0:t+1} = x^*_0:t+1$ and $N - 1$ independent samples from

$$w^*_t \delta_{x^*_0:t+1} (dx'_0:t+1) + \sum_{i=2}^N w^i_{t+1} \delta_{x^i_{0:t+1}} (dx'_0:t+1)$$
cPF example
cPF example
Uniform Ergodicity

Assumption

- \( g_\theta(x_t, y_t) \leq G_{t,\theta} \) (density upper bounded)
- Predicted density lower bounded:

\[
\int m_\theta(dx_0)g_\theta(x_0, y_0) \geq \frac{1}{G_{0,\theta}}, \quad \int m_\theta(x_{t-1}, dx_t)g_\theta(x_t, y_t) \geq \frac{1}{G_{t,\theta}}
\]

For any \( \theta \) and \( \varepsilon > 0 \), for \( N \) large enough,

\[
|(P_{N,\theta}\varphi)(x_{0:T}) - (P_{N,\theta}\varphi)(\tilde{x}_{0:T})| \leq \varepsilon
\]

for all \( x_{0:T}, \tilde{x}_{0:T} \), and \( \varphi : \mathcal{X}^{T+1} \to [-1, 1] \)

\( \Rightarrow \) Large \( N \) gives samples arbitrarily close to target \( p(x_{0:T}|\theta, y_{0:T}) \)

\( \Rightarrow \) Uniform ergodicity
Use coupling: define \((X^*_{0:T}, \dot{X}^*_{0:T})\) such that

\[
\text{Law}(X^*_{0:T}) = P_{N,\theta}(x_{0:T}, dx^*_{0:T}) \quad \text{Law}(\dot{X}^*_{0:T}) = P_{N,\theta}(\dot{x}_{0:T}, d\dot{x}^*_{0:T})
\]

If

\[
P(X^*_{0:T} \neq \dot{X}^*_{0:T}) \leq \epsilon
\]

then

\[
P_N^T(x_{0:T}, A) - P_N^T(\dot{x}_{0:T}, A) = \mathbb{E} \left\{ \left( \mathbb{I}_A(X^*_{0:T}) - \mathbb{I}_A(\dot{X}^*_{0:T}) \right) \mathbb{I}_{\{X_{0:T} \neq \dot{X}_{0:T}\}} \right\}
\leq P(X^*_{0:T} \neq \dot{X}^*_{0:T})
\leq \epsilon
\]
**Aim**: Coupling outputs of $P_{N,\theta}(x_{0:T}, dx_{0:T}^*)$ and $P_{N,\theta}(\tilde{x}_{0:T}, d\tilde{x}_{0:T}^*)$

- If cPF outputs $\{x_{0:t}^i\}_{i=1}^N$ and $\{\tilde{x}_{0:t}^i\}_{i=1}^N$ satisfy

  $x_{0:t}^i = \tilde{x}_{0:t}^i$ for $i \in C_t$

  where $C_0 = \{2, \ldots, N\}$

- Same holds after **append** move: $(x_{0:t}^i, x_{t+1}^i) = (\tilde{x}_{0:t}^i, \tilde{x}_{t+1}^i)$ for $i \in C_t$

- Resampling move: select particles in $C_t$ for survival if possible

**Coupling probability determined by law of $C_T$**
N. Whiteley (RSS discussion of PMCMC) suggested an extra \textit{backward} step for cPF that tries to modify (recursively, backward in time) the ancestry of the selected trajectory.

There is a forward “version” by Lindsten and Schon (2012)
Idea of Backward Sampling
Left Statistic: counting proportion $x'_i \neq x_i$ for $i = 0, \ldots, 399$
Success depends on state transition law $m_\theta(x_t, dx_{t+1})$

The cPF kernel $P_{\theta,N}$ with a BS step dominates the no BS version in asymptotic efficiency

i.e. gives rise to a CLT with a smaller asymptotic variance (Idea: self-adjoint operator + lag-1 domination; see Tierney (1998), Mira & Geyer (1999))

The cPF kernel geometrically ergodic $\Rightarrow$ cPF kernel with BS is geometrically ergodic

(Idea: both kernels are positive operators)
Established uniform ergodicity of cPF using coupling

Dependance on $N$ and $T$ not given although in practise $N$ should scale linearly with $T$ – now proved by (Andrieu, Lee, Vihola) and (Douc, Lindsten, Moulines)

Whiteley’s BS is better: in asymptotic efficiency and inherits geometric ergodicity